

АНАЛИЗ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ СООБЩЕНИЙ В СОЦИАЛЬНЫХ СЕТЯХ (НА ПРИМЕРЕ СЕТИ «ВКОНТАКТЕ»)

И. Е. Воронина, В. А. Гончаров

Воронежский государственный университет

Поступила в редакцию 13.11.2015 г.

Аннотация. Работа посвящена исследованию действующих методов анализа эмоциональной окраски текстов, разработке и реализации методики определения тональности текста, с помощью которой на основе выборки данных из социальной сети «ВКонтакте» анализируется тональность введенного пользователем сообщения.

Ключевые слова: анализ тональности, машинное обучение, наивный байесовский классификатор, социальные сети.

Annotation. This work is devoted to research of existing methods of sentiment analysis of texts, to developing and implementing methodology of sentiment analysis with which, based on a data sets from the social media network «VKontakte», analyzes the tone of message entered by user.

Keywords: sentiment analysis, machine learning, naive Bayes classifier, social networks.

ВВЕДЕНИЕ

Задача анализа эмоциональной окраски текстов, развитие методов фильтрации в сети Интернет приобретают всё большую актуальность в связи с огромной аудиторией сети, растущим средним временем нахождения в ней, а также большим охватом среди детей и подростков. Аналитика и мониторинг социальных сетей представляет огромный интерес для социологов, лингвистов, психологов, маркетологов и государственных структур.

Для решения задач анализа эмоциональной окраски текста в компьютерной лингвистике используются методы контент-анализа, общее название для которых – Sentiment Analysis (анализ тональности текста).

Крупные компании используют социальные сети для исследования мнений о своих продуктах. Такой подход, в отличие от использования опросов на сайте производителя

и работы фокус-групп, обеспечивает большую широту исследования мнений.

В органах, обеспечивающих государственную безопасность, контент-анализ используется для фильтрации и выявления сообщений, содержащих информацию о противоправных действиях (террористические угрозы, анализ запрещенного контента на сайтах и др.).

В социологических исследованиях методы контент-анализа позволяют не только показать текущее отношение населения к объекту исследования, но и прогнозировать изменения отношения к объекту. Например, в США существует проект Pulse of the Nation [1], целью которого является определение настроения граждан страны в течение дня путем исследования и анализа их записей в популярной социальной сети Twitter.

Определение тональности текстовых сообщений способно выявить эмоционально окрашенную лексику и проанализировать оценку автора по отношению к объектам, речь о которых идет в тексте.

1. ПОСТАНОВКА ЗАДАЧИ

Учитывая актуальность задачи систему анализа эмоциональной окраски сообщений, были разработаны программные средства, позволяющие однозначно определять тональность введенных пользователем данных. Анализ проводился на основе выборки массива информации из социальной сети «ВКонтакте».

Для разработки и реализации программных средств решались следующие задачи:

- анализ существующих методов анализа решений по оценке эмоциональной окраски текстов;

- исследование особенностей написания текстовых сообщений в социальной сети с целью выбора необходимого для анализа метода;

- разработка метода автоматизированного анализа эмоционально окрашенных сообщений в социальной сети «ВКонтакте» по определенной выборке данных.

Требования к системе анализа эмоциональной окраски сообщений:

1. Выбор метода анализа тональности должен быть адекватен специфике и структуре сообщений социальной сети.

2. Необходимо определять эмоциональную окраску текста на основе выборки сообщений социальной сети «ВКонтакте» по трем категориям: позитивная, нейтральная и негативная.

3. Точность определения тональности не менее 70 %.

4. Система анализа должна быть реализована как web-приложение и быть доступна с любого устройства, имеющего выход в сеть Интернет.

При реализации были использованы следующие технологии и программные средства: API VK (предоставляет методы для взаимодействия с сервисом «ВКонтакте»); Microsoft Visual Studio; Microsoft Azure Machine Learning (решение от Майкрософт, объединяющее мощност комплексного машинного обучения и преимущества облачных технологий).

2. АНАЛИЗ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ ТЕКСТА

В [2] под анализом эмоциональной окраски текста понимается задача автоматического анализа эмоционально окрашенной лексики и мнений, выраженных в тексте. При анализе эмоциональной окраски полагают, что текстовая информация в сети Интернет и, в частности, в социальных сетях подразделяется на два класса: мнения и факты. Наиболее значимым понятием является определение мнения. Мнения делятся на два типа: простое мнение и сравнение.

Простое мнение содержит точку зрения автора о каком-либо одном объекте. Может быть высказано прямо или неявно.

При анализе тональности текста выделяется 3 типа эмоций: позитивная, негативная и нейтральная. Текст считается нейтрально окрашенным если он не содержит в себе эмоциональной составляющей.

Сравнение – процесс количественного или качественного сопоставления разных свойств (сходств, отличий, преимуществ и недостатков) двух или более объектов с целью формирования определенной оценки или мнения для одного из них [2]. Выделяются 3 типа сравнений: превосходство одного объекта над другими (Superlative); сравнение аспектов объектов в пользу одного (Nonequal Gradable); приравнивание аспектов разных объектов (Equative).

Для анализа эмоциональной окраски текста необходимо алгоритмически разрешить проблемы извлечения аспектов из эмоционально окрашенного мнения и определения полярности мнения или текста.

3. МЕТОДЫ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ АСПЕКТОВ

Метод извлечения аспектов из мнения направлен на идентификацию объектов, относительно которых высказано какое-либо мнение.

Наиболее распространенные методы для извлечения аспектов – это метод обучения без учителя и статистический метод.

Выбор именно этих методов обусловлен их независимостью от размеченных тренировочных данных, которые отсутствуют в свободном доступе.

3.1. Метод обучения без учителя

Раздел машинного обучения, в котором определяются закономерности и взаимосвязи между объектами из некоторой неразмеченной выборки данных, называется обучением без учителя. Противоположным в реализации этому методу является метод обучения с учителем, в котором для каждого объекта из обучающей выборки задается правильный ответ и требуется найти взаимосвязь между объектами и ответами.

Рассмотрим метод распространения (bootstrapping) применительно к задаче извлечения аспектов. С помощью небольшого числа вручную определенных примеров определенного класса, возможно определять и автоматически извлекать необходимые текстовые единицы, постепенно наращивая их множество.

Шаг 1. Задать множество $U_0 = \{s_1, \dots, s_k\}$, $i = 0$.

Шаг 2. Найти в документе n -граммы $n_1, \dots, n_m$, близкие к терминам из U .

Шаг 3. $U_{i+1} = U_i + \{n_1, \dots, n_m\}$, $i = i + 1$.

Шаг 4. Перейти к шагу 2, если $i < n$.

Здесь U_0 – исходное множество терминов, n -граммы – слова или их части, состоящие из n символов, U_i – новое множество терминов, с включенными проанализированными n -граммами.

Аспектами могут быть только существительные, прилагательные, наречия и глаголы, представленные в виде n -грамм размера 1–3 слова. Схожесть аспектов из множества U_i и n -граммы определяется с помощью метрики $R \log F$ (1).

$$R \log (ng) = \log (ng, U_i) \times \frac{\text{freq}(ng, U_i)}{\text{freq}(ng)}. \quad (1)$$

Здесь $\text{freq}(ng, U)$ – частота встречаемости терминов из U и n -граммы ng во фрагменте текста, состоящего из определенного количества слов.

3.2. Статистические методы

В [1] авторы полагают, что термины, которые описывают аспекты какого-либо объекта, могут быть представлены либо одиночными существительными, либо словосочетаниями, содержащими в своем составе существительное, которые наиболее часто встречаются при анализе мнений объектов одного типа. Частота более одного процента считается достаточной для выбора удовлетворяющих этим требованиям n -грамм слов из общей массы терминов.

Выбранные n -граммы дополнительно проверяются на компактность, если они состоят из более чем двух слов.

Алгоритм определения компактности:

Шаг 1. F – n -грамма из n слов, а S – предложение, содержащее все n -граммы из F .

Шаг 2. F считается компактной в предложении S , если расстояние между двумя словами, которые смежны в F , в самом предложении S менее четырех слов.

Шаг 3. Если n -грамма проходит проверку на компактность хотя бы в двух предложениях, она заносится в список аспектов.

Также необходимо проверять термины, состоящие из одного слова.

В данном случае используется следующий алгоритм проверки:

Шаг 1. Выбираются все предложения, содержащие термин.

Шаг 2. Из них выбираются все предложения, не прошедшие тест на компактность для n -грамм, содержащих данный термин.

Шаг 3. Экспериментально определяется порог количества таких предложений, по значению которого можно судить, является ли выбранный термин аспектом.

При решении задач определения терминов-аспектов на практике часто используется метод C -value. Данный метод определяет значимость всех n -грамм, входящих в некоторое множество текстов по формуле (2).

$$C\text{-value} = \begin{cases} \log_2(\text{len}(\text{term})) \times \text{freq}(\text{term}), & \text{если } |e_{\text{terms}}| = 0; \\ \log_2(\text{len}(\text{term})) \times \text{freq}(\text{term}) - & \\ - \frac{1}{|e_{\text{terms}}|} \sum_{e_{\text{terms}}} \text{freq}(\text{elder}), & \text{иначе.} \end{cases} \quad (2)$$

Здесь *term* – *n*-грамма; e_{terms} – множество всех *n*-грамм большего порядка, содержащих *term*; $|e_{\text{terms}}|$ – мощность множества; *elder* – элемент данного множества; *len(term)* – длина.

4. МЕТОДЫ ОПРЕДЕЛЕНИЯ ПОЛЯРНОСТИ

Определить полярность текста, значит проанализировать его эмоциональную окраску и определить, положительно она или негативно. В случае если не удастся причислить текст или сообщение к одному из типов полярности, его считают нейтрально окрашенным.

Полярность текста определяется как на уровне предложения, так и в общем на уровне документа, текста, сообщения и т. д. Для определения полярности текста необходимо решить задачу определения полярности отдельных предложений.

Наиболее часто полярность определяется в два этапа: проверка на субъективность и непосредственно определение полярности. Объективное предложение выражает фактическую информацию о каком-либо объекте и не рассматривается с точки зрения определения полярности. Субъективное, напротив, чаще всего содержит мнение автора об объекте, поэтому оно наиболее пригодно для анализа эмоциональной окраски предложения и текста.

Текст, в отличие от предложения, часто может содержать части как с негативной, так и с позитивной тональностью. Поэтому усредненное значение полярности документа в современных системах анализа мнений считается неинформативно оценкой, не позволяющей достоверно утверждать, как именно эмоционально окрашен текст.

Классическим методом определения полярности мнения является словарный. Он основан на словарях эмоционально окрашенных слов и символов (смайлы, знаки препинания). Каждому термину из словаря присваивается свой вес, вручную, либо при помощи сторонних инструментов.

Существует несколько основных методов определения тональности с помощью словарей. Например, каждый термин словаря оценивается по двум направлениям: позитивному и негативному и по определенной шкале (например, от –10 до 10). Позитивная окраска сообщения или текста формируется исходя из веса термина с максимальной положительной окраской, входящего в текст. Также определяется и негативная окраска. В случае отсутствия термина в словаре, ему присваивается минимальный положительный вес по шкале.

Основным недостатком при использовании словарных методов является процедура составления словарей терминов и указания их весов, так как практически всегда значение веса термина зависит от выбранной предметной области. Например, если для грузового автомобиля термин «тяжелый» скорее является положительно окрашенным, то для гоночного болида он явно несет негативную окраску.

Применительно к социальным сетям, эффективным считается метод классификации. Сообщения в этом случае обладают некоторым рейтингом сообщения (например, количество «лайков» на записи в социальной сети), по которому и происходит классификация. Аспекты же извлекаются непосредственно из текста: выбираются *n*-граммы слов определенных частей речи и классифицируются при помощи различных методов машинного обучения, например, алгоритм обучения с учителем «наивный байесовский классификатор». Основным достоинством данного классификатора является малое количество исходных данных, требуемых для точной оценки параметров, необходимых для классификации.

Вероятностная модель классификатора основана на формуле Байеса.

Общая схема работы системы анализа тональности с использованием наивного байесовского классификатора:

1. Создание модели (включает в себя программирование модели для тренировки).
2. Обучение модели (создается тренировочный словарь с эмоционально окрашенными сообщениями и подготавливается тренировочный массив данных для обучения классификатора).
3. Использование модели (программирование расчета эмоциональной окраски).

5. АНАЛИЗ МЕТОДОВ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ

Задача определения полярности сообщений, текстов или предложений имеет множество эффективных методов решения, в каждом из которых есть особенности.

Словарные методы не позволяют создать универсальный словарь терминов, так как их вес в разных предметных областях может значительно отличаться или быть противоположным.

Методы обучения с учителем сложны составлением тренировочного словаря для предметной области, в которой используется классификатор, но в тоже время показывают наиболее высокую точность анализа.

Применительно к задаче извлечения аспектов, существуют различные методы фильтрации, позволяющие отделять и не рассматривать слова, не относящиеся к аспектам.

В табл. 1 представлена сравнительная характеристика наиболее распространенных методов анализа тональности.

На основании проведенного анализа выбран метод машинного обучения с учителем с использованием наивного байесовского классификатора. Выбор метода обусловлен, в первую очередь, высокими показателями точности при анализе коротких сообщений, что актуально для поставленной задачи. Дополнительно, данный метод позволяет работать с различными исходными массивами данных для обучения системы, что может значительно повлиять на точность анализа. В табл. 2 приведены примеры эмоционально окрашенных сообщений и соответствующие им результаты работы классификатора.

Данные для анализа взяты из социальной сети новостного портала «РИА-Воронеж» (<http://vk.com/riavrn>).

6. ОБУЧЕНИЕ МОДЕЛИ

Для реализации системы распознавания тональности использован метод обучения с учителем, а именно наивный байесовский классификатор. Суть метода состоит в обучении машинного классификатора на коллекции заранее размеченных текстов. Для обучения была составлена размеченная выборка, содержащая набор некоторых текстов («постов» социальной сети) и соответствующих им оценок окраски.

Таблица 1

Сравнение методов анализа тональности текстов

	Точность	Автоматизация	Данные для обучения	Простота применения	Применяемость в коммерческих системах
Подход на правилах	наиболее точный	подлежит	не требует данных	–	+
Подход со словарем	не универсален	в рамках одной предметной области	требует данные	+	–
Машинное обучение	точный	автоматический	требует данные	+/-	+
Обучение без учителя	низкая точность	автоматический	не требует данных	+	+

Таблица 2

Примеры работы системы анализа тональности

Сообщение	Тональность, определенная анализатором	Реальная тональность сообщения
Дико и страшно, что такое случается совсем близко: Воронежские разбойницы избили несовершеннолетнюю в День защиты детей. После инцидента в Отрожке девочка попала в больницу	Негативная	Негативная
Воронеж занял 12 место среди российских городов по популярности поездок с детьми летом 2015. Неплохой результат, если учесть, что моря у нас нет.	Негативная	Нейтральная/позитивная
Под Воронежем в подвале многоэтажки барсучиха родила детенышей (у вас, небось, только кошки живут? А в Рамони вон не только). Жильцы дома не особо довольны.	Негативная	Негативная
Стараемся быть скромными, но сегодня хвалимся: «Медиа-логия» снова признала РИА «Воронеж» самым цитируемым СМИ региона.	Позитивная	Позитивная
Сегодня в Воронеже под колесами Kia погибла косуля. Животное выскочило на дорогу на улице Новосибирской. Мы пообщались с очевидцем и ГИБДД	Негативная	Нейтральная
Воронежские врачи отказались делать аборт в День защиты детей.	Нейтральная	Нейтральная
Бомбардировщик Су-34 опрокинулся при посадке в Воронежской области. По предварительным данным, самолет из-за отказа тормозного парашюта выкатился за пределы взлетно-посадочной полосы.	Негативная	Негативная
РИА «Воронеж» и сеть кинотеатров СИНЕМА ПАРК представляют вам уникальную возможность выиграть пригласительный на закрытый предпоказ фантастического блокбастера.	Позитивная	Позитивная

Таблица 3

Выборка данных для обучения классификатора

Оценка	Сообщение
0	«Вальядолид» готов вести переговоры о трансфесе Эберта в «Спартак»
5	Не надо быть правильным, надо быть настоящим
-5	– так много есть нельзя
-5	– не сиди в интернете

Выборка состоит из более чем 3000 сообщений в новостной ленте социальной сети «ВКонтакте». Сообщения выбраны случайным образом исходя из даты их публикации. Выборка представлена в виде файла .csv следующего вида (табл. 3).

Каждому выбранному сообщению была проставлена оценка эмоциональной окраски в границах от -10 до 10. Выбор данного диа-

пазона позволяет наиболее точно оценить сообщения не усложняя задачу классификации.

Для выборки построена матрица документ-термин, в которой строкам соответствуют записи, а столбцам – встречающиеся в них термины. Ячейка таблицы содержит количество повторений конкретного термина в документе.

Таблица 4

Матрица-термин для записей из выборки

	сегодня	хороший	легкий	ветерок	я	не	очень	хорошо	погулял	из-за	ветра
Запись 1	1	1	1	1	0	0	1	0	0	0	0
Запись 2	1	0	0	0	1	1	1	1	1	1	1

Таблица 5

Тестирование работы программы

	Кол-во тестовых сообщений	Правильно проанализированных сообщений	Точность (%)
Сообщений с позитивной окраской	147	106	72
Сообщений с негативной окраской	153	103	68

Например, для двух записей «Сегодня очень хороший легкий ветерок» и «Я не очень хорошо сегодня погулял из-за ветра» матрица документ-термин будет выглядеть следующим образом (табл. 4).

Для того чтобы избежать присутствия среди терминов нескольких форм одного и того же слова (например в табл. 4 «хороший» и «хорошо», «ветерок» и «ветра») можно использовать стемминг – процесс поиска основы слова для выбранного исходного слова.

ЗАКЛЮЧЕНИЕ

Исследование методов контент-анализа эмоциональной окраски текстов применительно к процессу анализа мнений на примере данных из социальной сети «ВКонтакте» привело к реализации прототипа веб-сервиса для анализа тональности вводимых пользователем сообщений. Тональность сообщения определялась по трем категориям: позитивная, нейтральная и негативная.

Проведенное тестирование системы с применением технологии стемминга и без нее, с различными по длине и эмоциональной окраске сообщениями выявило, что точность работы алгоритма пропорциональна длине сообщения и составляет до 89 % при анализе сообщений длиной до 50 символов. Менее 70 % сообщений система анализирует верно при длине сообщений от 200 и более символов. Для тестирования точности определения эмоциональной окраски сообщений было

проведено более 300 запросов к источникам, содержащим как позитивно, так и негативно окрашенные сообщения. Результаты тестирования представлены в табл. 5.

Полученные результаты, безусловно, не решают проблему определения эмоциональной окраски текстов, так как задача определения тональности невероятно сложна, сильно зависима от предметной области и вследствие этого объемна. Однако, исследование проблем контент-анализа в социальных сетях может оказать влияние на совершенствование систем социологических исследований, фильтрации нежелательного контента, поиска сообщений, содержащих угрозы и т. д. Исследования на основе данных социальной сети «ВКонтакте» особенно интересны, так как сеть является крупнейшей в России и странах СНГ с аудиторией около 220 млн. пользователей [3]. Методы анализа социальной сети могут также стать шагом к созданию принципиально новых автоматизированных социологических и маркетинговых исследований тональности в конкретной предметной области.

СПИСОК ЛИТЕРАТУРЫ

1. Панг Б. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval / Б. Панг, Л. Ли. – Москва : Вильямс, 2008. – 235 с.

Воронина Ирина Евгеньена – профессор кафедры программного обеспечения и администрирования информационных систем факультета прикладной математики, информатики и механики, доктор технических наук.
E-mail: irina.voronina@gmail.com

Гончаров Виталий Андреевич – специалист по ведению маркетинговых исследований в CPA-сети Adcombo.
E-mail: goncharov.vitaly@gmail.com

2. Панг Б. Thumbs up? Sentiment Classification using Machine Learning Techniques / Б. Панг, Л. Ли. – Москва : Вильямс, 2002. – 312 с.

3. Социальные сети в России. – URL: <http://corp.mail.ru/> (дата обращения: 08.05.2015).

Voronina Irina E. – Professor of Software Development and Information Systems Administration Department of Applied Mathematics, Informatics and Mechanics Faculty, PhD, Dr. habil. in Technical Sciences.
E-mail: irina.voronina@gmail.com

Goncharov Vitaly Andreevich – Marketing research manager in the international CPA network Adcombo.
E-mail: goncharov.vitaly@gmail.com